BÀI GIẢNG THỐNG KÊ

Giảng viên
Vũ Đỗ Huy Cường

Khoa Toán-Tin học
Đại học Khoa học Tự nhiên
vdhuycuong@gmail.com

Mục lục

Chương 1

Mẫu và Đặc trưng

1.1.1 Biến và dữ liệu

Biến và dữ liệu

- Biến (Variable): một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang người hay vật, hiện tượng khác.
- Biến định tính (quanlitative variable): biểu diễn tính chất của đặc trưng nó thể hiện, có tác dụng phân loại; ví dụ : nhóm máu (A, B, AB, O), giới tính (nam, nữ), màu mắt (đen, nâu, xanh),. . .
- Biến định lượng(quantitative variable): biểu diễn độ lớn của đặc trưng mà nó thể hiện; ví dụ: chiều cao, cân nặng, thời gian, ...
  Biến định lượng: bao gồm biến rời rạc (discrete variable) và biến liên tục (continuous variable).
- Dữ liệu (data): các giá trị của một biến. Tập hợp tất cả những quan trắc cho một biến cụ thể được gọi là một tập dữ liệu (Data set).

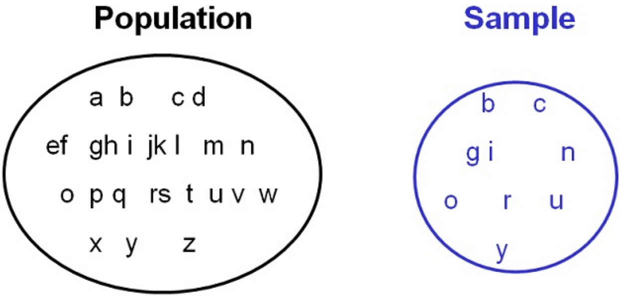
1.1.2 Tổng thể và mẫu

Tổng thể và mẫu

- Tổng thể (population) Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Mẫu (sample) là một tập con được chọn ra từ tổng thể. Ta thường ký hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.
- Tham số (parameter) là một đặc trưng cụ thể của một tổng thể.
- Thống kê(statistic) là một đặc trưng cụ thể của mẫu.

1.1.2 Tổng thể và mẫu

Tổng thể và mẫu



Những giá trị tính từ dữ liệu của tổng thể gọi là các tham số

Những giá trị được tính từ dữ liệu của mẫu gọi là các thống kê

1.1.2 Tổng thể và mẫu

Ví dụ 1.1.

a) Số cử tri đi bầu cử.

b) Điểm trung bình của tất cả các sinh viên trong một trường đại học.

c) Thu nhập của các hộ gia đình trong thành phố.

d) Trọng lượng các sản phẩm trong một nhà máy.

Thông thường, ta không thể chọn hết tất cả các phần tử của tổng thể để nghiên cứu bởi vì

- số phần tử của tổng thể rất lớn.
- thời gian và kinh phí không cho phép.
- có thể làm hư hại các phần tử của tổng thể

Do đó, ta chỉ thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

1.1.3 Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có tỉ lệ được chọn như nhau (xác suất được chọn bằng nhau).
- Mọi mẫu cỡ n có cùng tỉ lệ được chọn từ tổng thể.

Phương pháp chọn mẫu ngẫu nhiên đơn giản (simple random sampling)

- Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.
- Trộn đều các phiếu, sau đó chọn có hoàn lại n phiếu. Các phần tử của tổng thể có số thự tự trong phiếu lấy ra sẽ được chọn làm mẫu.

1.1.3 Chọn mẫu ngẫu nhiên

Thông thường, người ta sắp xếp mẫu theo dạng Bảng.

Ví dụ 1.2.

a) Kiểm tra điểm số ngẫu nhiên 50 sinh viên. Ta sắp xếp điểm số X và số sinh viên n có điểm tương ứng vào bảng như sau.

X	2	4	5	6	7	8	9	10
n	4	6	20	10	5	2	2	1

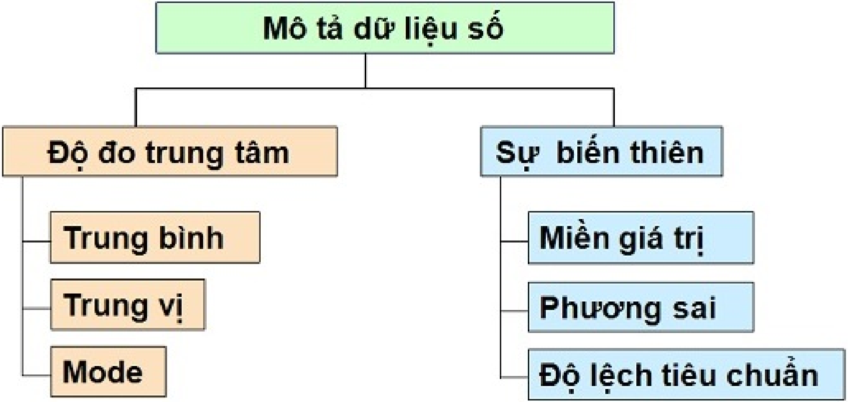
b) Đo chiều cao X (cm) của 100 thanh niên. Khi đó, ta có bảng số liệu ở dạng khoảng như sau:

X	148-152	152-156	156-160	160-164	164-168
n	5	20	35	25	15

Khi cần tính toán, người ta sử dụng bảng trung bình

X	150	154	158	162	166
n	5	20	35	25	15

1.2.1 Trung bình



1.2.1 Trung bình

Định nghĩa 1 (Trung bình(mean))

Nếu một tổng thể có N phần tử được kí hiệu là $X_1, X_2, \ldots, X_N$, thì trung bình tổng thể là

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

Nếu n quan sát của một mẫu được kí hiệu là $X_1, X_2, \ldots, X_n$, thì trung bình mẫu là

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Nhận xét 1

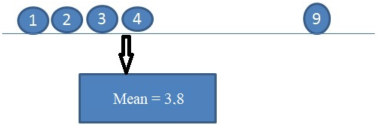
Trung bình : bị ảnh hưởng bởi các ngoại lai (outlier).

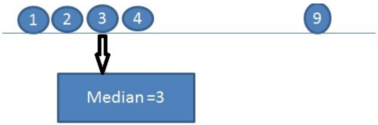
1.2.2 Trung vị

Định nghĩa 2 (Trung vị (median))

Trung vị (median) là giá trị chia các quan sát thành hai phần bằng nhau. Một phần chứa các quan sát nhỏ hơn trung vị và phần còn lại chứa các quan sát lớn hơn trung vị.

- *Trung vị không bị ảnh hưởng bởi các outlier.*
- *Vị trí của trung vị (sắp xếp dữ liệu tang dần):* $i = \frac{n+1}{2}$
    - *Nếu* $i \in \mathbb{Z}$ *trung vị* $= X_i$.
    - *Nếu* $i \notin \mathbb{Z}$ *trung vị* $= \frac{X_{[i]}+X_{[i]+1}}{2}$, *với [i] là phần nguyên của i.*



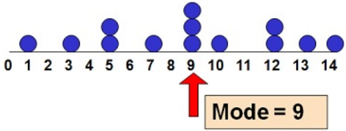


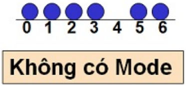
1.2.3 Tần số

Định nghĩa 3 (Tần số (Mode))

Mode là một đại lượng để đo xu hướng trung tâm của dữ liệu, nó là giá trị thường xảy ra nhất,

- *Không bị ảnh hưởng bởi các outlier*
- *Có thể có nhiều mode hoặc không tồn tại mode*

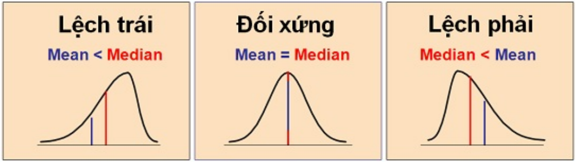




1.2.3 Tần số

Ta có thể xác định vị trí tương đối của trung bình, trung vị và mode từ đồ thị tổ chức tần số như sau:

- Nếu phân phối tần số là đối xứng, thì trung bình = trung vị =mode.
- Nếu phân phối tần số bị lệch (skewed) (tức là bất đối xứng, với một đuôi dài về một phía), thì trung bình và trung vị đều bị kéo về phía đuôi dài hơn, nhưng trung bình, thông thường được kéo xa hơn trung vị. Cụ thể, nếu phân phối là lệch phải thì mode< trung vị < trung bình; ngược lại, nếu phân phối là lệch trái thì mode> trung vị > trung bình.



1.2.3 Tần số

Ví dụ 1.3.
Tìm Trung bình (mean), Trung vị (median), Tần số (mode) của dữ liệu sau

X	1,2	2,0	3,5	5,9	7,2	9,1
n	1	1	2	3	2	3

Giải
Trung bình: Mean

$$= \frac{1,2 + 2,0 + 3,5 \cdot 2 + 5,9 \cdot 3 + 7,2 \cdot 2 + 9,1 \cdot 3}{12} = 5,8.$$

Trung vị: Vị trí $= \dfrac{12+1}{2} = 6,5$. Median $= \dfrac{x_6 + x_7}{2} = \dfrac{5,9+5,9}{2} = 5,9.$

Tần số: Mode $= 5,9$ và $9,1$.

1.2.3 Tần số

Bài tập

1.1. Tìm Trung bình (mean), Trung vị (median), Tần số (mode) của dữ liệu sau

X	2,2	2,6	3,1	4,0	5,3	5,7
n	1	2	1	2	2	3

1.2. Tìm Trung bình (mean), Trung vị (median), Tần số (mode) của dữ liệu sau

X	1,1	1,6	2,5	3,1	5,6	8,1	9,2
n	2	1	1	3	2	3	3

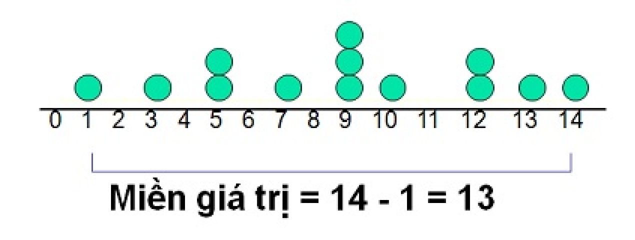
1.3.1 Miền giá trị

Định nghĩa 4 (Miền giá trị (range))

- *Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.*
- *Là độ lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc.*

Nếu n quan sát trong một mẫu được kí hiệu $x_1, x_2, \ldots, x_n$ thì miền giá trị mẫu là

$$r = max(x_i) - min(x_i)$$



1.3.2 Phương sai - Độ lệch chuẩn

Định nghĩa 5 (Phương sai)

Phương sai tổng thể

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Với N là số phần tử của tổng thể.
Độ lệch chuẩn tổng thể là $\sigma = \sqrt{\sigma^2}$.

Phương sai mẫu gồm n quan trắc là

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Độ lệch chuẩn mẫu là $S = \sqrt{S^2}$.

1.3.2 Phương sai - Độ lệch chuẩn

Định lý 1

Với một tổng thể bất kì có trung bình μ và độ lệch chuẩn σ, và $k > 1$, thì phần tram các giá trị quan trắc nằm trong khoảng

$$[\mu - k\sigma, \mu + k\sigma]$$

bằng ít nhất $100\left[1 - \frac{1}{k^2}\right]\%$.

Ví dụ 1.4.

ít nhất	nằm trong
$(1 - 1/1^2) = 0\%$	$(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$(\mu \pm 3\sigma)$

1.3.3 Đặc trưng của mẫu

Xét một mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$, ta có các đặc trưng sau:

- Trung bình mẫu: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
- Phương sai mẫu: $\hat{S}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

Phương sai mẫu hiệu chỉnh $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

- Tỉ lệ mẫu: Xét mẫu định tính với các biến X_i có phân phối Bernoulli $B(1, p)$.

$$X_i = \begin{cases} 0 & \text{nếu phần tử không có tính chất A} \\ 1 & \text{nếu phần tử có tính chất A} \end{cases}$$

Nếu mẫu có m phần tử có tính chất A thì tỉ lệ mẫu là

$$F = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{m}{n}$$

1.3.3 Đặc trưng của mẫu

Ví dụ 1.5.

Tìm trung bình mẫu, phương sai mẫu, phương sai hiệu chỉnh mẫu và tỉ lệ mẫu thỏa $X > 4$

X	1	2	3	5	6	8	10

Giải

Trung bình mẫu: $\bar{x} = \dfrac{1}{7}(1+2+3+5+6+8+10) = 5.$

Phương sai mẫu:

$$\hat{s}^2 = \frac{1}{7}\Big((1-5)^2 + (2-5)^2 + ... + (10-5)^2\Big) = 9,1429.$$

PSM hiệu chỉnh:

$$s^2 = \frac{1}{7-1}\Big((1-5)^2 + (2-5)^2 + ... + (10-5)^2\Big) = 10,6667.$$

Tỉ lệ mẫu thỏa $X > 4$: $f = \dfrac{4}{7} = 0,5714$

1.3.3 Đặc trưng của mẫu

Bài tập

1.3. Tìm trung bình mẫu, phương sai mẫu, phương sai hiệu chỉnh mẫu và tỉ lệ mẫu thỏa $X > 4$

X	-2	-1	0	2	3	5	7

1.4. Tìm trung bình mẫu, phương sai mẫu, phương sai hiệu chỉnh mẫu và tỉ lệ mẫu thỏa X chẵn

X	1	2	4	5	6	8	10	12

1.4.1 Thống kê của mẫu

Định nghĩa 6 (Thống kê (statistic))

Một thống kê (statistic) là một hàm bất kì các quan sát trong một mẫu ngẫu nhiên.

- Trung bình mẫu: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Phương sai mẫu: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$
- Độ lệch chuẩn mẫu: $S = \sqrt{S^2}$
- Giá trị nhỏ nhất của mẫu: $Y_1 = \min(X_1, X_2, \ldots, X_n)$
- Giá trị lớn nhất của mẫu: $Y_2 = \max(X_1, X_2, \ldots, X_n)$
- $R = Y_n - Y_1$

đều là các thống kê của mẫu ngẫu nhiên có kích thước n.

1.4.2 Phân phối mẫu của trung bình và phương sai

Định lý 2

Giả sử $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên được lấy từ một tổng thể có phân phối chuẩn với trung bình μ và phương sai σ^2. Khi đó,

- $\bar{X}$ *và* S^2 *độc lập nhau.*
- $\bar{X} \sim N(\mu, \sigma^2/n)$ *với N là phân phối chuẩn.*
- $\frac{(n-1)S^2}{\sigma} \sim \chi^2(n-1)$ *với* χ *là phân phối Chi;*
- $\frac{\bar{X}-\mu}{S\sqrt{n}} \sim T(n-1)$ *với T là phân phối Student;*

Trong thực hành khi mẫu có kích thước đủ lớn ($n \geq 30$), ta có các phân phối xấp xỉ chuẩn sau:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1) \qquad \frac{\bar{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$$

1.4.3 Phân phối mẫu của tỉ lệ

Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1 & \text{nếu thỏa } \mathcal{A} \\ 0 & \text{nếu khác} \end{cases}$$

thu được mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$,với $X_i \sim B(1, p)$ trong đó p là tỉ lệ phần tử thỏa đặc trưng $\mathcal{A}$.
Đặt $X = \sum_{i=1}^{n} X_i$ là số phần tử thỏa đặc trưng $\mathcal{A}$ trong mẫu khảo sát, thì $X \sim B(n, p)$.
Tỉ lệ mẫu $\hat{p}$ là một ước lượng của tỉ lệ p được xác định bởi

$$\hat{p} = \frac{X}{n}$$

1.4.3 Phân phối mẫu của tỉ lệ

Kì vọng và phương sai của $\hat{p}$ là

$$\mathbb{E}(\hat{p}) = p, \ \mathbb{V}ar = \frac{p(1-p)}{n}$$

Theo định lí giới hạn trung tâm ta có

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1).$$

Trong thực hành, khi $np \geq 5, \ n(1-p) \geq 5$ thì

$$\hat{p} \approx N(p, \frac{p(1-p)}{n}).$$

Chương 2

Ước lượng điểm

2.1.1 Giới thiệu

Giả sử cần khảo sát một đặc tính X thuộc một tổng thể xác định. Biến ngẫu nhiên X có phân phối $F(\theta)$, trong đó tham số θ chưa biết và $\theta \in \Theta$ một tập hợp tham số.

Bài toán: tìm tham số θ.

Chọn mẫu ngẫu nhiên cỡ n từ $X: \; X_1, X_2, \ldots, X_n$.

Thống kê $\hat{\theta} = h(X_1, \ldots, X_n)$ gọi là một ước lượng điểm (Point estimator) cho θ.

Với một mẫu thực nghiệm $x_1, \ldots, x_n$, ta gọi $\hat{\theta} = h(x_1, \ldots, x_n)$ là một giá trị ước lượng điểm (point estimate) cho θ.

2.1.1 Giới thiệu

Ví dụ 2.1.
Đặt $X =$ Chiều cao dân số trong một khu vực, $X \sim N(\mu, \sigma^2)$.
Phân phối của X phụ thuộc vào kỳ vọng μ và phương sai σ^2.
Thống kê trung bình mẫu và phương sai mẫu.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

là những ước lượng điểm cho μ và σ^2.
Với một mẫu thực nghiệm $x_1 = 150, x_2 = 155, x_3 = 167$, giá trị ước lượng điểm của μ và σ^2 là $\bar{x} = 157.333$, $s^2 = 76.333$.

2.1.2 Ước lượng không chệch

Định nghĩa 7 (Ước lượng không chệch)

Ước lượng điểm $\hat{\theta}$ gọi là một ước lượng không chệch (Unbiased estimator) cho tham số θ nếu: $E(\hat{\theta}) = \theta$

Nếu $\hat{\theta}$ là ước lượng chệch của θ, độ sai khác $E(\hat{\theta}) - \theta$ gọi là độ chệch của ước lượng, ký hiệu $Bias(\hat{\theta})$.

Ví dụ 2.2.

a) $\bar{X}$ là một ước lượng không chệch của μ

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

b) S^2 là một ước lượng không chệch của σ^2.

c) $\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ là ước lượng chệch của σ^2 vì $E(\hat{S}^2) \neq \sigma^2$.

2.1.3 Ước lượng hiệu quả - bền vững

Định nghĩa 8 (Ước lượng hiệu quả)

Xét $\hat{\theta}$ và $\bar{\theta}$ là hai ước lượng không chệch của θ, $\hat{\theta}$ gọi là ước lượng hiệu quả hơn $\bar{\theta}$ nếu với một cỡ mẫu n cho trước

$$Var(\hat{\theta}) < Var(\bar{\theta})$$

Định nghĩa 9 (Ước lượng vững)

Gọi $\hat{\theta}_n = h(X_1, \ldots, X_n)$ là một ước lượng điểm của θ. Ước lượng $\hat{\theta}_n$ được gọi là ước lượng vững nếu với mọi $\varepsilon > 0$

$$\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| \leq \varepsilon\right) = 1$$

Ví dụ 2.3.

a) S^2 là ước lượng vững của σ^2.

b) Với $X \sim B(n, p)$, $\hat{p} = \frac{\bar{X}}{n}$ là ước lượng vững cho p.

2.2.1 Ước lượng trung bình

Giả sử X là BNN với $EX = \mu$ (chưa biết). μ được gọi là giá trị trung bình của tập hợp chính.
Nếu ta có một mẫu n gồm giá trị $x_1, x_2, \ldots, x_n$ của X thì trung bình mẫu

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

sẽ được dùng làm ước lượng cho μ.

Định lý 3

Trung bình mẫu là ước lượng không chệch và vững cho trung bình của tập hợp chính.

2.3.1 Ước lượng phương sai

Giả sử X là BNN với $Var(X) = \sigma^2$ (chưa biết). σ^2 được gọi là phương sai của tập hợp chính. Nếu ta có một mẫu gồm n giá trị $x_1, x_2, \dots, x_n$ của X thì một cách hợp lí phương sai mẫu chưa hiệu chỉnh

$$\widehat{S}^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}$$

được xem xét dùng để ước lượng σ^2. Tuy nhiên phương sai mẫu chưa hiệu chỉnh $\widehat{s}^2$ là một ước lượng chệch.

Do đó, nếu ta xét phương sai mẫu đã hiệu chỉnh

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

thì S^2 là một ước lượng không chệch của σ^2.

2.4.1 Ước lượng tỉ lệ

Giả sử ta quan tâm đến một đặc tính A nào đó mà mỗi cá thể của tập hợp chính Ω có thể mang hoặc không mang. Gọi p là tỉ lệ cá thể mang đặc tính A trong Ω. Chúng muốn ước lượng p dựa trên việc khảo sát một mẫu gồm n cá thể.

Xét biến lượng X xác định như sau:

$$X = \begin{cases} 0 & \text{nếu cá thể không có đặc tính } A \\ 1 & \text{nếu cá thể có đặc tính } A \end{cases}$$

Từ định nghĩa của X ta có

- $P(X = 0) = 1 - p$.
- $P(X = 1) = p$.

2.4.1 Ước lượng tỉ lệ

Nếu $X_1, X_2, \ldots, X_n$ là một mẫu gồm n giá trị quan sát của X thì $X_1 + X_2 + \ldots + X_n$ là số cá thể mang đặc tính A của mẫu và

$$f = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

chính là tần suất xuất hiện đặc tính A trong mẫu.

Ta thấy f là giá trị quan sát của BNN

$$F = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

trong đó $X_1, X_2, \ldots, X_n$ là các BNN độc lập với nhau và có cùng phân bố với X. Vì $EX = p$ nên ta dễ dàng chứng minh được f là một ước lượng không chệch và vững cho p.

Chương 3

Ước lượng khoảng

3.1.1 Khoảng ước lượng

Cho biến ngẫu nhiên X có phân phối $F(\theta)$. Giả sử cần khảo sát một đặc tính X trên một tổng thể xác định. Chọn một mẫu ngẫu nhiên cỡ $n: \ X = (X_1, \ldots, X_n)$.

Định nghĩa 10 (Ước lượng khoảng (Interval estimator))

Một ước lượng khoảng của một tham số θ là một cặp các thống kê $L(X_1, \ldots, X_n)$. và $U(X_1, \ldots, X_n)$ của một mẫu ngẫu nhiên thỏa $L(X) \leq U(X)$ và $L(X) \leq \theta \leq U(X)$.

Nếu một mẫu thực nghiệm $x = (x_1, \ldots, x_n)$ được quan trắc, $[l(x), u(x)]$ gọi là một khoảng ước lượng (interval estimate) cho θ.

3.1.2 Khoảng tin cậy

Định nghĩa 11 (Khoảng tin cậy)

Xét biến ngẫu nhiên $X = (X_1, \ldots, X_n)$ *là biến ngẫu nhiên có hàm mật độ đồng thời phụ thuộc vào tham số* $\theta \in \mathbb{R}$ *và* $L(X), U(X)$ *là hai thống kê sao cho* $L(X) \leq U(X)$.

Khi đó, khoảng ngẫu nhiên $[L(X), U(X)]$ *gọi là khoảng tin cậy cho tham số* θ *với độ tin cậy* $100(1-\alpha)\%$ *nếu*

$$P\{L(X) \leq \theta \leq U(X)\} = 1 - \alpha$$

Ví dụ 3.1.

Với mẫu thực nghiệm $x = (x_1, \ldots, x_n)$, ta có khoảng tin cậy cho tham số θ là $l(x) \leq \theta \leq u(x)$ với độ tin cậy 95%

Ý nghĩa: với 200 lần lấy mẫu cỡ n thì

có 190 lần giá trị tham số $\theta \in [l, u]$;

có 10 lần giá trị tham số $\theta \notin [l, u]$.

3.2.1 Ước lượng khoảng cho trung bình - Biết phương sai

Các giả định:

- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 của tổng thể đã biết.

Xây dựng khoảng tin cậy:

- Tìm trung bình mẫu $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Tra bảng PP CHUẨN tìm t_α tương ứng với $\varphi(t_\alpha) = \dfrac{1-\alpha}{2}$.
- Khoảng ước lượng là

$$(\bar{x} - \epsilon, \bar{x} + \epsilon) \text{ với } \epsilon = t_\alpha \frac{\sigma}{\sqrt{n}}$$

3.2.2 ULKTB - Không biết phương sai, mẫu lớn

Các giả định

- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 của tổng thể không biết
- Cỡ mẫu lớn: $n > 30$.

Xây dựng khoảng tin cậy:

- Tìm $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ và $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Tra bảng PP CHUẨN tìm t_α tương ứng với $\varphi(t_\alpha) = \dfrac{1-\alpha}{2}$.
- Khoảng ước lượng là

$$(\bar{x} - \epsilon, \bar{x} + \epsilon) \text{ với } \epsilon = t_\alpha \frac{s}{\sqrt{n}}$$

3.2.3 ULKTB - Không biết phương sai, mẫu nhỏ

Các giả định

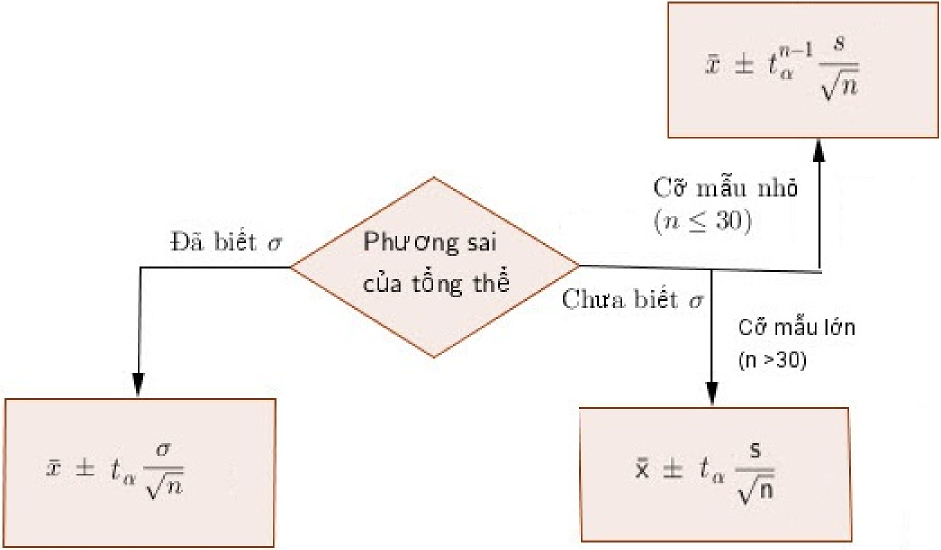
- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 của tổng thể không biết
- Cỡ mẫu nhỏ: $n \leq 30$.

Xây dựng khoảng tin cậy:

- Tìm $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ và $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Tra bảng PP STUDENT tìm t_{α}^{n-1} tương ứng với α và $n-1$.
- Khoảng ước lượng là

$$(\bar{x} - \epsilon, \bar{x} + \epsilon) \text{ với } \epsilon = t_{\alpha}^{n-1}\frac{s}{\sqrt{n}}$$

3.2.4 Tổng hợp các trường hợp



3.2.4 Tổng hợp các trường hợp

Ví dụ 3.2.

Đo chiều cao (đơn vị cm) 100 sinh viên trường Đại học Tôn Đức Thắng ta được trung bình mẫu $\overline{x} = 160$ (cm). Giả sử độ lệch chuẩn σ của chiều cao người trưởng thành là 8 (cm).

Hãy xác định khoảng ước lượng chiều cao trung bình của sinh viên trường ĐHCN với độ tin cậy 95%.

Giải

Từ đề bài ta có $n = 100; \overline{x} = 160; \sigma = 8$.
Với độ tin cậy 95% ta được $\alpha = 0,05$. Tra BẢNG PHÂN PHỐI CHUẨN, ta thu được $t_\alpha = 1,96$.
Vậy khoảng ước lượng chiều cao trung bình của sinh viên TĐTU là

$$\left(\overline{x} - t_\alpha \frac{\sigma}{\sqrt{n}}; \overline{x} + t_\alpha \frac{\sigma}{\sqrt{n}}\right) = (158,43; 161,57).$$

3.2.4 Tổng hợp các trường hợp

Ví dụ 3.3.

Để xác định chiều cao trung bình (đơn vị m) của các cây bạch đàn trong một khu rừng bạch đàn rất lớn, người ta chọn ngẫu nhiên 64 cây để đo. Kết quả thu được như sau:

Chiều cao	5,5-6,5	6,5-7,5	7,5-8,5	8,5-9,5	9,5-10,5
Số cây	6	15	20	13	10

Với độ tin cậy 98%, hãy xác định khoảng ước lượng chiều cao trung bình của cây bạch đàn trong khu rừng.

Giải

Để dễ tính toán, mỗi khoảng chiều cao ta sẽ lấy trung điểm của khoảng làm đại diện. Từ đây, ta tính được $n = 64; \overline{x} = 8,09; s = 1,2$.
Với độ tin cậy 95% ta được $\alpha = 0,05$. Tra BẢNG PHÂN PHỐI CHUẨN, ta thu được $t_\alpha = 2,33$.
Vậy khoảng ước lượng trung bình chiều cao của cây bạch đằng trong khu rừng là $\left(\overline{x} - t_\alpha \frac{s}{\sqrt{n}}; \overline{x} + t_\alpha \frac{s}{\sqrt{n}}\right) = (7,74; 8,44)$.

3.2.4 Tổng hợp các trường hợp

Ví dụ 3.4.

Một chỉ tiêu để đánh giá hiệu quả của phương pháp trị bệnh là số ngày trung bình μ từ lúc điều trị cho đến khi bệnh nhân khỏi bệnh. Một mẫu ngẫu nhiên gồm 16 bệnh nhân được theo dõi và số ngày điều trị cho tới khi bệnh nhân khỏi bệnh được ghi lại như sau:

4	4	5	8	6	10	3	9
2	6	4	7	9	11	6	8

Với độ tin cậy 95%, hãy xác định khoảng ước lượng trung bình số ngày cần thiết để bệnh nhân được điều trị hết bệnh.

Giải

Từ đề bài ta tính được $n = 16; \overline{x} = 6,375; s = 2,630$.

Với độ tin cậy 95% ta được $\alpha = 0,05$. Tra BẢNG PHÂN PHỐI STUDENT, ta thu được $t_\alpha^{15} = 2,131$. Khi đó, KULTB ... là

$$\left(\overline{x} - t_\alpha^{n-1}\frac{s}{\sqrt{n}}; \overline{x} + t_\alpha^{n-1}\frac{s}{\sqrt{n}}\right) = (4,974; 7,776).$$

3.2.4 Tổng hợp các trường hợp

Bài tập

3.1. Đường kính của một ống piston trong động cơ xe máy có phân phối chuẩn với độ lệch chuẩn $\sigma = 0.001$ mm. Một mẫu ngẫu nhiên gồm 15 ống piston có đường kính trung bình $\bar{x} = 74.036$ mm.

a) Lập KTC 90% cho đường kính trung bình của piston.
b) Lập KTC 99% cho đường kính trung bình của piston.

3.2. Đo chỉ số IQ của các sinh viên trong 1 trường đại học, khảo sát 18 sinh viên thu được kết quả sau:

130	122	119	142	136	127	120	152	141
132	127	118	150	141	133	137	129	142

Biết rằng chỉ số IQ của sinh viên theo phân phối chuẩn với $\sigma = 10,50$.

a) Lập khoảng tin cậy 95% cho chỉ số IQ trung bình.
b) Lập khoảng tin cậy 99% cho chỉ số IQ trung bình.

3.2.4 Tổng hợp các trường hợp

3.3. CAdmiun (Cd), một kim loại nặng, là chất độc đối với các loài động vật. Tuy nhiên , các loài nấm lại có khả nang hấp thụ Cd với hàm lượng cao. Chính phủ một số nước ra quy định giới hạn hàm lượng Cd tối đa trong rau quả khô là 0.5 (ppm). Trong một số nghiên cứu về hàm lượng Cd trong loài nấm Boletus pinicola của M.Melgar và các cộng sự trên tạp chí Journal of Environment Science and Health, cho số liệu về hàm lượng Cd trong một mẫu gồm 12 cây nấm như sau:

0.24	0.59	0.62	0.16	0.77	1.33
0.92	0.19	0.33	0.25	0.59	0.32

a) Lập khoảng tin cậy 99% cho hàm lượng Cd trung bình trong nấm Boletus pinicola.

b) Nếu muốn sai số ước lượng $\varepsilon = 0.1$ thì khảo sát tối thiểu bao nhiêu cây nấm?

3.2.4 Tổng hợp các trường hợp

3.4. Biết tháng lương (Đv: Triệu đồng) của thanh niên trong độ tuổi 25-35 ở một khu vực có phân phối chuẩn. Khảo sát 50 thanh nhiên.

Lương tháng	1.8	2.5	3.2	3.9	4.6	5.3	6.0	6.7	7.4
Số thanh niên	2	3	8	9	11	7	6	2	2

a) Lập khoảng tin cậy 95% cho lương tháng của thanh niên trong khu vực này.

b) Nếu muốn sai số ước lượng $\varepsilon = 0.10$ mà vẫn giữ cỡ mẫu $n = 50$ thì độ tin cậy còn bao nhiêu?

3.3.1 Ước lượng khoảng cho tỉ lệ

Các giả định

- Lấy mẫu ngẫu nhiên cỡ n. Đặt $Y =$ số phần tử thỏa tính chất A trong n phần tử khảo sát, thì Y có phân phối Bernoulli

Xây dựng khoảng tin cậy:

- Tìm $\bar{f} = \dfrac{m}{n}$ (m là số phần tử tương ứng của Y) và $s^2 = \dfrac{\bar{f}(1-\bar{f})}{n}$
- Tra bảng PP CHUẨN tìm t_α tương ứng với $\varphi(t_\alpha) = \dfrac{1-\alpha}{2}$.
- Khoảng ước lượng là

$$(\bar{f} - \epsilon, \bar{f} + \epsilon) \text{ với } \epsilon = t_\alpha s$$

3.3.1 Ước lượng khoảng cho tỉ lệ

Ví dụ 3.5.

Trước ngày bầu cử tổng thống, một cuộc tham dò dư luận đã được tiến hành. Người ta chọn ngẫu nhiên 400 người để hỏi ý kiến thì có 240 người nói rằng họ sẽ bỏ phiếu cho ông *A*. Tìm khoảng ước lượng cho tỉ lệ cử tri bỏ phiếu cho ông *A* với độ tin cậy 95%.

Giải

Từ đề bài ta tính được

$$n = 400; f = \frac{240}{400} = 0,6, s^2 = \sqrt{\frac{f(1-f)}{n}} = 0,0245.$$

Với đô tin cậy 95% ta được $\alpha = 0,05$. Tra BẢNG PHÂN PHỐI CHUẨN, ta thu được $t_\alpha = 1,96$. Vì

$$\begin{cases} nf = 400 \times 0,6 = 240 > 10 \\ n(1-f) = 400 \times 0,4 = 160 > 10 \end{cases}$$

nên khoảng ước lượng tỉ lệ cử tri bỏ phiếu cho ông *A* là

$$(f - t_\alpha s; f + t_\alpha s) = (55,2\%; 64,8\%).$$

3.3.1 Ước lượng khoảng cho tỉ lệ

Bài tập

3.5. Biết tháng lương (Đv: Triệu đồng) của thanh niên trong độ tuổi 25-35 ở một khu vực có phân phối chuẩn. Khảo sát 50 thanh nhiên.

Lương tháng	1.8	2.5	3.2	3.9	4.6	5.3	6.0	6.7	7.4
Số thanh niên	2	3	8	9	11	7	6	2	2

a) Thanh niên có thu nhập cao nếu lương tháng từ 6.0 triệu đồng trở lên, Hãy lập khoảng tin cậy 95% cho tỷ lệ có thu nhập cao.

b) Nếu muốn sai số ước lượng $\varepsilon = 0.08$ thì cần khảo sát thêm bao nhiêu thanh niên nữa?

3.6. Theo dõi 1000 bệnh nhân ung thư phổi thấy có 823 bệnh nhân chết trong vòng 10 nam.

a) Lập KTC 95% cho tỷ lệ bệnh nhân chết vì ung thư phổi.

b) Nếu muốn sai số bé hơn 0.03 thì phải theo dõi tối thiểu bao nhiêu bệnh nhân trong 10 nam?

3.4.1 Khoảng tin cậy cho phương sai

Định lý 4

Giả sử $X_1, X_2, \ldots, X_n$ là n quan sát từ BNN $X \sim N(\mu, \sigma^2)$. Khi đó, $T = \frac{(n-1)S^2}{\sigma^2}$ có phân bố χ^2 với bậc tự do $n-1$, tức $T \sim \chi^2(n-1)$ và

$$S^2 = \frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1}$$

là phương sai đã hiệu chỉnh của mẫu $X_1, X_2, \ldots, X_n$.

Ta gọi λ^2_α là giá trị sao cho $P(T > \lambda^2_\alpha) = \alpha$. Khi đó,

$$P\left(\lambda^2_{1-\frac{\alpha}{2}} < T < \lambda^2_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Vậy, khoảng ước lượng cho σ^2 với độ tin cậy $1-\alpha$ là

$$\left(\frac{(n-1)\,S^2}{\lambda^2_{\frac{\alpha}{2}}}; \frac{(n-1)\,S^2}{\lambda^2_{1-\frac{\alpha}{2}}}\right).$$

3.4.1 Khoảng tin cậy cho phương sai

Ví dụ 3.6.

Đường kính của một chi tiết máy do Xí nghiệp A sản xuất là một BNN có phân phối chuẩn. Chọn ngẫu nhiên 9 chi tiết máy ta tính được độ lệch chuẩn đã hiệu chỉnh là $s = 0,1$. Với độ tin cậy 95%, xác định khoảng ước lượng cho phương sai của toàn bộ chi tiết máy do Xí nghiệp A sản xuất.

Giải

Từ giả thiết đề bài ta có

$$n = 9; s = 0,1; \lambda^2_{\frac{\alpha}{2}} = 17,535; \lambda^2_{1-\frac{\alpha}{2}} = 2,18.$$

Khoảng ước lượng cho phương sai của toàn bộ chi tiết máy do Xí nghiệp A sản xuất là $\left(\frac{(n-1)S^2}{\lambda^2_{\frac{\alpha}{2}}}; \frac{(n-1)S^2}{\lambda^2_{1-\frac{\alpha}{2}}}\right) = (0,0046; 0,0367)$.

3.4.1 Khoảng tin cậy cho phương sai

Bài tập

3.7. . Khảo sát 16 sinh viên về điểm trung bình của học kỳ 2 thì tính được s = 1,5 điểm. Hãy ước lượng phương sai về điểm trung bình học kỳ 2 của sinh viên với độ tin cậy 97%, biết rằng điểm trung bình X của sinh viên là biến ngẫu nhiên có phân phối chuẩn.

3.8. . Mức hao phí nguyên liệu cho 1 đơn vị sản phẩm là biến ngẫu nhiên X (gram) có phân phối chuẩn. Quan sát 28 sản phẩm này người ta thu được bảng số liệu:

X (gram)	19,0	19,5	20	20,5
Số sản phẩm	5	6	14	3

Với độ tin cậy 95%, hãy ước lượng phương sai của mức hao phí nguyên liệu trên trong 2 trường hợp:

a) biết EX = 20 gram;

b) chưa biết EX.

Chương 4

Kiểm định tham số

4.1.1 Giả thuyết và kiểm định

Ví dụ 4.1.

Giám đốc một nhà máy sản xuất bo mạch chủ máy vi tính tuyên bố rằng tuổi thọ trung bình của một bo mạch chủ do nhà máy sản xuất là 60 tháng; đây là một giả thuyết về kỳ vọng của biến ngẫu nhiên $X =$ tuổi thọ của một bo mạch chủ. Để đưa ra kết luận là chấp nhận hay bác bỏ giả thuyết trên, ta cần dựa vào mẫu điều tra và quy tắc kiểm định thống kê.

Định nghĩa

- Giả thuyết thống kê là những phát biểu về các tham số, quy luật phân phối, hoặc tính độc lập của các đại lượng ngẫu nhiên.
- Kiểm định giả thuyết là quá trình mà qua đó có thể quyết định bác bỏ giả thuyết hay không, dựa vào mẫu $(X_1, \ldots, X_n)$ lấy từ tổng thể.

4.1.2 Giả thuyết không và đối thuyết

Định nghĩa

Trong bài toán kiểm định giả thuyết,

- Giả thuyết không (null hypothesis) là giả thuyết cần được kiểm định. Ký hiệu: H_0.
- Đối thuyết (alternative hypothesis) là giả thuyết được xem xét để thay thế giả thuyết không. Ký hiệu: H_1.

Xét bài toán kiểm định tham số, giả sử ta quan trắc mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ từ biến ngẫu nhiên X có hàm mật độ xác suất $f(x;\theta)$ phụ thuộc vào tham số θ. Gọi Θ là không gian tham số, và Θ_0 và Θ_0^c là hai tập con rời nhau của Θ sao cho $\Theta_0 \cup \Theta_0^c = \Theta$. Giả thuyết (giả thuyết không) và đối thuyết của bài toán có dạng như sau

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_0^c \end{cases}$$

4.1.2 Giả thuyết không và đối thuyết

Ví dụ 4.2.

Gọi μ độ thay đổi trung bình trong huyết áp của một bệnh nhân sau khi dùng thuốc bác sĩ điều trị cần quan tâm đến giả thuyết sau

$H_0 : \mu = 0$	Không có ảnh hưởng thuốc lên huyết áp của bệnh nhân
$H_1 : \mu \neq 0$	Có ảnh hưởng của thuốc lên huyết áp của bệnh nhân.

Ví dụ 4.3.

Một khách hàng quan tâm đến tỷ lệ sản phẩm kém chất lượng trong một lô hàng mua của một nhà cung cấp. Giả sử tỷ lệ sản phẩm kém tối đa được phép là 5%. Khách hàng cần quan tâm đến giả thuyết sau

$H_0 : p \geq 0.05$	Tỷ lệ sản phẩm kém cao hơn mức cho phép
$H_1 : p < 0.05$	Tỷ lệ sản phẩm kém ở mức chấp nhận được

4.1.2 Giả thuyết không và đối thuyết

Tổng quát, một bài toán kiểm định giả thuyết cho tham số θ sẽ có một trong 3 dạng dưới đây (θ_0 là giá trị kiểm định đã biết): Hai phía:

$$\begin{cases} H_0 & : \theta = \theta_0 \\ H_1 & : \theta \neq \theta_0 \end{cases}$$

Một phía bên trái:

$$\begin{cases} H_0 & : \theta \geq \theta_0 \\ H_1 & : \theta < \theta_0 \end{cases}$$

Một phía bên phải:

$$\begin{cases} H_0 & : \theta \leq \theta_0 \\ H_1 & : \theta > \theta_0 \end{cases}$$

4.1.3 Miền bác bỏ - Tiêu chuẩn kiểm định

Định nghĩa

Xét bài toán kiểm định giả thuyết H_0 và đối thuyết H_1. Giả sử rằng H_0 đúng, từ mẫu ngẫu nhiên $X = (X_1, \ldots, X_n)$ chọn hàm $Z = h(X_1, \ldots, X_n; \theta_0)$ sao cho với số $\alpha > 0$ bé tùy ý ta có thể tìm dược tập hợp W_α thỏa điều kiện

$$P(Z \in W_\alpha) = \alpha$$

Tập hợp W_α gọi là miền bác bỏ giả thuyết H_0 và phần bù W_α^c gọi là miền chấp nhận giả thuyết H_0. Đại lượng $Z = h(X_1, \ldots, X_n; \theta_0)$ gọi là tiêu chuẩn kiểm định giả thuyết H_0. Giá trị α gọi là mức ý nghĩa của bài toán kiểm định.

Nếu $z \in W_\alpha$ thì ta bác bỏ giả thuyết H_0.

Nếu $z \in W_\alpha^c$ thì ta chưa đủ cơ sở để bác bỏ H_0 (chấp nhận h_0).

4.1.3 Miền bác bỏ - Tiêu chuẩn kiểm định

Bài toán Kiểm định

- Để giải quyết bài toán kiểm định, ta quan sát mẫu ngẫu nhiên $X_1, ..., X_n$ và đưa ra giả thuyết H.
- Từ mẫu trên, ta chọn thống kê $T = F(X_1, ..., X_n, \theta)$ sao cho nếu khi H đúng thì phân phối xác suất của T hoàn toàn xác định.
- Với mức ý nghĩa α, ta tìm được khoảng tin cậy (hay khoảng ước lượng) $[a, b]$ cho T ở độ tin cậy $1 - \alpha$. Khi đó:
  - nếu $t \in [a, b]$ thì ta chấp nhận giá thuyết H.
  - nếu $t \notin [a, b]$ thì ta bác bỏ giá thuyết H.

4.1.4 Sai lầm loại I và loại II

Trong bài toán kiểm định, ta có thể mắc phải các sai lầm sau

a. Sai lầm loại I: là sai lầm mắc phải khi ta bác bỏ H_0 trong khi thực tế giả thuyết H_0 đúng. Xác suất mắc sai lầm loại I: $\alpha = P(W_\alpha | H_0)$

b. Sai lầm loại II: là sai lầm mắc phải khi ta chấp nhận giả thuyết H_0 trong khi thực tế H_0 sai. Xác suất mắc sai lầm: $\beta = P(W_\alpha^c | H_1)$

Trong thống kê, người ta quy ước rằng sai lầm loại 1 tác hại hơn loại 2 nên cần tránh hơn.

Quyết định \ Thực tế	H_0 đúng	H_0 sai
Không bác bỏ H_0	Không có sai lầm $(1-\alpha)$	Sai lầm loại II β
Bác bỏ H_0	Sai lầm loại I α	Không có sai lầm $(1-\beta)$

4.2.1 Kiểm định giả thuyết cho kỳ vọng - Biết Phương sai

Các giả định

- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 đã biết, kì vọng μ chưa biết

Kiểm định: so sánh μ với μ_0 cho trước

1. Phát biểu giả thuyết không H_0.
2. Tìm trung bình mẫu $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
3. Tính thống kê kiểm định $t_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
4. Từ mức ý nghĩa α tra bảng PP CHUẨN tìm

$$t_\alpha \text{ thỏa } \varphi(t_\alpha) = \frac{1-\alpha}{2} \qquad \text{hoặc } z_\alpha \text{ thỏa } \varphi(z_\alpha) = 0.05 - \alpha.$$

4.2.1 Kiểm định giả thuyết cho kỳ vọng - Biết Phương sai

Kiểm định: so sánh μ với μ_0 cho trước

5. Xác định miền bác bỏ W_α : bảng bên dưới

Giả Thuyết	Miền bác bỏ
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$W_\alpha = \{t_0 : \lvert t_0 \rvert > t_\alpha\}$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$W_\alpha = \{t_0 : t_0 > z_\alpha\}$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$W_\alpha = \{t_0 : t_0 < -z_\alpha\}$

6. Kết luận: Bác bỏ H_0 (Chưa đủ cơ sở bác bỏ H_0).

4.2.2 KĐGT cho kỳ vọng - Không biết Phương sai, mẫu lớn

Các giả định

- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 chưa biết, kì vọng μ chưa biết, mẫu lớn $n > 30$

Kiểm định: so sánh μ với μ_0 cho trước

1. Phát biểu giả thuyết không H_0.
2. Tìm $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ và $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
3. Tính thống kê kiểm định $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Từ mức ý nghĩa α tra bảng PP CHUẨN tìm

$$t_\alpha \text{ thỏa } \varphi(t_\alpha) = \frac{1-\alpha}{2} \qquad \text{hoặc} \qquad z_\alpha \text{ thỏa } \varphi(z_\alpha) = 0.05 - \alpha.$$

4.2.2 KĐGT cho kỳ vọng - Không biết Phương sai, mẫu lớn

Kiểm định: so sánh μ với μ_0 cho trước

5. Xác định miền bác bỏ W_α : bảng bên dưới

Giả Thuyết	Miền bác bỏ
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$W_\alpha = \{t_0 : \lvert t_0\rvert > t_\alpha\}$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$W_\alpha = \{t_0 : t_0 > z_\alpha\}$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$W_\alpha = \{t_0 : t_0 < -z_\alpha\}$

6. Kết luận: Bác bỏ H_0 (Chưa đủ cơ sở bác bỏ H_0).

4.2.3 KĐGT cho kỳ vọng - Không biết Phương sai, mẫu nhỏ

Các giả định

- Mẫu ngẫu nhiên được chọn từ tổng thể có phân phối chuẩn.
- Phương sai σ^2 chưa biết, kì vọng μ chưa biết, Mẫu nhỏ $n \leq 30$

Kiểm định: so sánh μ với μ_0 cho trước

1. Phát biểu giả thuyết không H_0.
2. Tìm $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ và $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
3. Tính thống kê kiểm định $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Từ mức ý nghĩa α tra bảng PP STUDENT tìm
   t_{α}^{n-1} hoặc $z_{2\alpha}^{n-1}$

4.2.3 KĐGT cho kỳ vọng - Không biết Phương sai, mẫu nhỏ

Kiểm định: so sánh μ với μ_0 cho trước

5. Xác định miền bác bỏ W_α : bảng bên dưới

Giả Thuyết	Miền bác bỏ
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$W_\alpha = \{t_0 : \lvert t_0 \rvert > t_\alpha^{n-1}\}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$W_\alpha = \left\{t_0 : t_0 > z_{2\alpha}^{n-1}\right\}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$W_\alpha = \left\{t_0 : t_0 < -z_{2\alpha}^{n-1}\right\}$

6. Kết luận: Bác bỏ H_0/ Chưa đủ cơ sở bác bỏ H_0.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Ví dụ 4.4.

Những thống kê trong nam 2008 cho thấy một người Mỹ đi du lịch ở Châu Âu trong vòng 3 tuần sẽ chi hết trung bình 1010 USD. Một cuộc nghiên cứu được tiến hành trong nam 2009 để xác định xem có sự thay đổi gì trong việc chi tiêu mua sắm khi du lịch Châu Âu của người Mỹ hay không. Khảo sát 100 khách du lịch cho thấy số tiền trung bình họ tiêu là 1015 USD. Hãy kiểm định giả thiết: “Số tiền trung bình một người Mỹ chi tiêu khi đi du lịch ở Châu Âu trong hai nam 2008 và 2009 là như nhau” với mức ý nghĩa 5%, biết độ lệch chuẩn qua từng nam là như nhau, bằng 300 USD.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Giải

Phát biểu giả thuyết H_0 và đối thuyết H_1:

$$H_0 : \mu = 1010$$
$$H_1 : \mu \neq 1010$$

với μ là số tiền trung bình một người Mỹ chi tiêu khi đi du lịch ở Châu Âu trong nam 2009.

Theo đề bài ta có $\overline{x} = 1015; \mu_0 = 1010; n = 100; \sigma = 300$.

Thống kê kiểm định được chọn là

$$t = \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sigma} = \frac{1}{6} = 0,1667$$

Tra BẢNG PHÂN PHỐI CHUẨN, ta thu được

$$\varphi(t_\alpha) = \frac{1 - 0.05}{2} = 0,475 \qquad \Rightarrow t_\alpha = 1,96.$$

Ta thấy $|t| < t_\alpha$ nên giả thiết H_0 được chấp nhận.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Ví dụ 4.5.

Một nghiên cứu cho rằng trung bình một khách hàng vào siêu thị A tiêu ít nhất 200 ngàn đồng. Ta muốn kiểm định khẳng định trên bằng cách chọn ngẫu nhiên 64 khách hàng. Với mẫu đã chọn, ta tính được số tiền trung bình họ tiêu là 220 ngàn đồng với độ lệch tiêu chuẩn là 50 ngàn đồng. Phát biểu giả thiết H_0, đối thiết H_1 và kiểm định H_0 với mức ý nghĩa $\alpha = 5\%$.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Giải

Trước hết, ta phát biểu giả thiết H_0 và đối thiết H_1:

$$H_0 : \mu = 200$$
$$H_1 : \mu > 200$$

với μ là số tiền trung bình của khách hàng chi tiêu trong siêu thị A.

Theo đề bài ta có $\overline{x} = 220$; $\mu_0 = 200$; $n = 64$; $s = 50$.

Thống kê kiểm định được chọn là

$$t = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s} = 3,2$$

Tra BẢNG PHÂN PHỐI CHUẨN, ta thu được

$\varphi(z_\alpha) = 0.05 - 0.5 = 0.45 \quad \Rightarrow z_\alpha = 1,645.$

Vì $t > z_\alpha$ nên ta bác bỏ H_0 và chấp nhận H_1.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Ví dụ 4.6.
Một công ty sản xuất pin tuyên bố rằng pin của họ có tuổi thọ trung bình là $21,5$ giờ. Một cơ quan kiểm tra chất lượng của 6 pin và thu được số liệu về tuổi thọ của 6 pin này (đơn vị, giờ):

$$19 \quad 18 \quad 22 \quad 20 \quad 16 \quad 25$$

Kết quả này có xác nhận là quảng cáo của công ty là đúng hay không? Mức ý nghĩa được chọn là $\alpha = 5\%$.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Giải

Trước hết, ta phát biểu giả thiết H_0 và đối thiết H_1 như sau:

$$H_0 : \mu = 21,5$$
$$H_1 : \mu \neq 21,5$$

với μ là tuổi thọ trung bình của pin trong thực tế.

Theo đề bài ta có $\overline{x} = 20$; $\mu_0 = 21,5$; $n = 6$; $s = 3,16$.

Thống kê kiểm định được chọn là

$$t = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s} = 1,16$$

Tra BẢNG PHÂN PHỐI STUDENT với mức ý nghĩa $\alpha = 0,05$ và bậc tự do $n - 1 = 5$, ta thu được $t_\alpha^5 = 2,571$.

Vì $|t| < t_\alpha^5$ nên ta chưa có cơ sở bác bỏ H_0.

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

Bài tập

4.1. Dây chuyền sản xuất kem đánh rang P/S được thiết kế để đóng hộp những tuýt kem có trọng lượng trung bình là 6 oz (1 oz = 28g). Một mẫu gồm 30 tuýt kem được chọn ngẫu nhiên để kiểm tra định kỳ. Bộ phận điều khiển dây chuyền phải đảm bảo để trọng lượng trung bình mỗi tuýt kem là 6 oz; nếu nhiều hơn hoặc ít hơn, dây chuyền phải được điều chỉnh lại.

Giả sử trung bình mẫu của 30 tuýt kem là 6.1 oz và độ lệch tiêu chuẩn của tổng thể $\sigma = 0.2$ oz.

Thực hiện kiểm định giả thuyết với mức ý nghĩa 3% để xác định xem dây chuyền sản xuất có vận hành tốt hay không?

4.2.4 Ví dụ và bài tập KĐGT cho kỳ vọng

4.2. Điểm trung bình môn Toán của sinh viên khóa trước là 5, 72. Khóa này, theo dõi 100 SV được số liệu:

Điểm	4	5	6	7	8
Số lượng	15	32	30	13	10

Kiểm định giả thuyết H : "điểm trung bình môn Toán của sinh viên khóa nay bằng khóa trước", mức ý nghĩa tối đa để *H* được chấp nhận là:

a)13,94%;

b)11,74%;

4.3. Thời gian X (phút) giữa hai chuyến xe bus trong một thành phố là biến ngẫu nhiên có phân phối chuẩn. Công ty xe bus nói rằng: trung bình cứ 5 phút lại có 1 chuyến xe bus. Người ta chọn ngẫu nhiên 8 thời điểm và ghi lại thời gian (phút) giữa hai chuyến xe bus là:

$$5,3; 4,5; 4,8; 5,1; 4,3; 4,8; 4,9; 4,7.$$

Với mức ý nghĩa 5%, hãy kiểm định lời nói trên ?

4.2.5 Kiểm định giả thuyết cho tỉ lệ

Các giả định

- Lấy mẫu ngẫu nhiên cỡ n. Đặt $Y =$ số phần tử thỏa tính chất A trong n phần tử khảo sát, thì Y có phân phối Bernoulli

Kiểm định: so sánh p với p_0 cho trước

1. Phát biểu giả thuyết không H_0.
2. Tìm $\bar{f} = \dfrac{m}{n}$ (m là số phần tử tương ứng của Y).
3. Tính thống kê kiểm định $t_0 = \dfrac{|\bar{f} - p_0|\sqrt{n}}{\sqrt{p_0(1-p_0)}}$
4. Từ mức ý nghĩa α tra bảng PP CHUẨN tìm

$$t_\alpha \text{ thỏa } \varphi(t_\alpha) = \frac{1-\alpha}{2} \qquad \text{hoặc} \qquad z_\alpha \text{ thỏa } \varphi(z_\alpha) = \frac{1}{2} - \alpha.$$

4.2.5 Kiểm định giả thuyết cho tỉ lệ

Kiểm định: so sánh p với p_0 cho trước

5. Xác định miền bác bỏ W_α : bảng bên dưới

Giả Thuyết	Miền bác bỏ
$H_0: \ p = p_0$ $H_1: \ p \neq p_0$	$W_\alpha = \{t_0 : \lvert t_0\rvert > t_\alpha\}$
$H_0: \ p = p_0$ $H_1: \ p > p_0$	$W_\alpha = \{t_0 : t_0 > z_\alpha\}$
$H_0: \ p = p_0$ $H_1: \ p < p_0$	$W_\alpha = \{t_0 : t_0 < -z_\alpha\}$

6. Kết luận: Bác bỏ H_0/ Chưa đủ cơ sở bác bỏ H_0.

4.2.5 Kiểm định giả thuyết cho tỉ lệ

Ví dụ 4.7.

Ở địa phương A người ta dự đoán có 45% số hộ gia đình sử dụng máy vi tính. Kiểm tra ngẫu nhiên 200 hộ ta thấy có 80 hộ dùng máy vi tính. Với mức ý nghĩa 5%, ta có kết luận dự đoán trên là chính xác?

Giải

Giả thuyết và đối thuyết: $H_0 : p = 45\%$ $\quad H_1 : p \neq 45\%$. với p là tỉ lệ hộ gia đình sử dụng máy vi tính.

Theo đề bài ta có $f = \frac{80}{200} = 0,4$; $p_0 = 0,45$; $n = 200$. Ta tính được

$$t = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1 - p_0)}} = -1,42$$

Tra bảng PHÂN PHỐI CHUẨN ta thu được

$$\varphi(t_\alpha) = \frac{1 - 0.05}{2} = 0,475 \qquad \Rightarrow t_\alpha = 1,96.$$

Vì $|t| < t_\alpha$ nên ta chấp nhận giả thiết H_0.

4.3.1 So sánh hai kỳ vọng - Biết phương sai

- Các giả định:
  - $X_1, \ldots, X_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_1 và phương sai σ_1^2.
  - $Y_1, \ldots, Y_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_2 và phương sai σ_2^2.
  - Tổng thể 1 và 2 (đại diện bởi X và Y) độc lập với nhau.
  - Các phương sai σ_1^2 và σ_2^2 đã biết.
- Bài toán kiểm định giả thuyết trên hai mẫu độc lập với mức ý nghĩa α cho trước gồm các dạng sau:

$$(a)\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 \neq D_0 \end{cases} \quad (b)\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 < D_0 \end{cases} \quad (c)\begin{cases} H_0 : \mu_1 - \mu_2 = D_0 \\ H_1 : \mu_1 - \mu_2 > D_0 \end{cases}$$

4.3.1 So sánh hai kỳ vọng - Biết phương sai

Các bước kiểm định

1. Phát biểu giả thuyết không và đối thuyết,
2. Xác định mức ý nghĩa α,
3. Tính thống kê kiểm định

$$z_0 = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}} \tag{1}$$

   nếu H_0 đúng, $Z_0 \sim N(0,1)$.
4. Xác định miền bác bỏ

Đối thuyết	**Miền bác bỏ**	**$p-$giá trị**
$H_1 : \mu_1 - \mu_2 \neq D_0$	$\lvert z_0 \rvert > z_{1-\alpha/2}$	$p = 2[1 - \Phi(\lvert z_0 \rvert)]$
$H_1 : \mu_1 - \mu_2 < D_0$	$z_0 < -z_{1-\alpha}$	$p = \Phi(z_0)$
$H_1 : \mu_1 - \mu_2 > D_0$	$z_0 > z_{1-\alpha}$	$p = 1 - \Phi(z_0)$

5. Kết luận: Nếu bác bỏ H_0, ta kết luận H_1 đúng với $(1-\alpha)100\%$ độ

4.3.1 So sánh hai kỳ vọng - Biết phương sai

Ví dụ 4.8.

Một công ty sản xuất sơn nghiên cứu về 1 loại phụ gia làm giảm thời gian khô của sơn. Thực hiên thí nghiệm trên 2 mẫu: mẫu thứ nhất gồm 10 vật mẫu được sơn bằng loại sơn bình thường; mẫu thứ hai gồm 10 mẫu vật được sơn với sơn có chất phụ gia mới. Trong những nghiên cứu trước, biết rằng độ lệch tiêu chuẩn của thời gian khô sau khi quét sơn là 8 phút và không thay đổi khi thêm phụ gia vào. Trung bình của mẫu 1 và mẫu 2 lần lượt là $\bar{x} = 121$ phút và $\bar{y} = 112$ phút. Với mức ý nghĩa 5%, hãy cho kết luận về loại sơn với chất phụ gia mới.

4.3.2 SS2KV - Không biết phương sai, mẫu lớn

- Các giả định:
  - $X_1, \ldots, X_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_1 và phương sai σ_1^2.
  - $Y_1, \ldots, Y_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_2 và phương sai σ_2^2.
  - Tổng thể 1 và 2 (đại diện bởi X và Y) độc lập với nhau.
  - Các phương sai σ_1^2 và σ_2^2 chưa biết.
  - Cỡ mẫu lớn: $n > 30$ và $m > 30$.

$$z_0 = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}} \tag{2}$$

  nếu H_0 đúng, $Z_0 \sim N(0, 1)$.
  - Miền bác bỏ (hoặc $p-$giá trị) trong trường hợp này đươc tính tương tụ trường hợp biết phương sai (thay thế σ_1 và σ_2 bởi S_1 và S_2.

4.3.2 SS2KV - Không biết phương sai, mẫu lớn

Ví dụ 4.9.

Khảo sát về chiều cao của sinh hai khoa Toán và CNTT: chọn ngẫu nhiên 50 sinh viên khoa Toán, tính được chiều cao trung bình là 163 cm và độ lệch tiêu chuẩn 5 cm. Đo chiều cao 50 khoa CNTT, có trung bình mẫu là 166 cm và độ lệch tiêu chuẩn 8 cm. Với mức ý nghĩa $\alpha = 1\%$, hãy cho kết luận về chiều cao của sinh viên hai khoa.

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

- Các giả định:
  - $X_1, \ldots, X_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_1 và phương sai σ_1^2.
  - $Y_1, \ldots, Y_n$ là mẫu ngẫu nhiên được chọn từ tổng thể 1 có phân phối chuẩn với kỳ vọng μ_2 và phương sai σ_2^2.
  - Tổng thể 1 và 2 (đại diện bởi X và Y) độc lập với nhau.
  - Các phương sai σ_1^2 và σ_2^2 chưa biết.
  - Cỡ mẫu nhỏ : $n \leq 30$ và $m \leq 30$.
- Ta xét hai trường hợp:
  1. Trường hợp phương sai bằng nhau $\sigma_1^2 = \sigma_2^2$,
  2. Trường hợp phương sai khác nhau $\sigma_1^2 \neq \sigma_2^2$.

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

- Giả sử $X_1, \ldots, X_n$ và $Y_1, \ldots, Y_m$ lần lượt là hai mẫu ngẫu nhiên chọn từ hai tổng thể độc lập và có phân phối chuẩn với kỳ vọng và phương sai là (μ_1, σ_1^2) và (μ_2, σ_2^2). Ta cần kiểm định giả thuyết

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

- Nếu S_1^2 là phương sai mẫu ngẫu nhiên $(X_1, \ldots, X_n)$ thì

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1)$$

tương tự, ta có

$$\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$$

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

- Khi đó, đại lượng

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

  sẽ có phân phối $\mathfrak{F}$ với $(n-1, m-1)$ bậc tự do.

- Xét biến ngẫu nhiên $F \sim \mathfrak{F}(u, v)$ có hàm mật độ xác suất là $f(x)$, phân vị trên mức α của F là $f_{\alpha,u,v)}$ được định nghĩa như sau

$$\mathbb{P}(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)dx = \alpha$$

- Phân vị mức $1-\alpha$ của F cho bởi

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}}$$

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

Các bước kiểm định

1. Phát biểu giả thuyết $H_0 : \sigma_1^2 = \sigma_2^2$ và đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$
2. Xác định mức ý nghĩa α
3. Khi H_0 đúng, thống kê
$$f = \frac{S_1^2}{S_2^2}$$
có phân phối $\mathfrak{F}$ với $(n-1, m-1)$ bậc tự do.
4. Xác định miền bác bỏ: Bác bỏ H_0 khi $|f| > f_{1-\alpha/2,n-1,m-1}$
5. Kết luận: Nếu bác bỏ H_0, ta kết luận H_1 đúng với $100(1-\alpha)\%$ độ tin cậy. Ngược lại kết luận chưa đủ cơ sở để bác bỏ H_0.

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

TH $\sigma_1^2 = \sigma_2^2 = \sigma^2$

- Ta sử dụng một ước lượng chung cho cả σ_1^2 và σ_2^2 là S_p^2 phương sai mẫu chung

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

- Thống kê

$$t_0 = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

có phân phối Student với $n+m-2$ bậc tự do.

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

TH $\sigma_1^2 = \sigma_2^2 = \sigma^2$

- Đặt $df = n + m - 2$, miền bác bỏ và $p-$ giá trị trong trường hợp này có dạng

Đối thuyết	**Miền bác bỏ**	$p-$**giá trị**
$H_1 : \mu_1 - \mu_2 \neq D_0$	$\|t_0\| > t^{df}_{1-\alpha/2}$	$p = 2\mathbb{P}(t^{df} \geq \|t_0\|)$
$H_1 : \mu_1 - \mu_2 < D_0$	$t_0 < -t^{df}_{1-\alpha}$	$p = \mathbb{P}(t^{df} \leq t_0)$
$H_1 : \mu_1 - \mu_2 > D_0$	$t_0 > t^{df}_{1-\alpha}$	$p = \mathbb{P}(t^{df} \geq t_0)$

- Kết luận: Bác bỏ H_0/ Chưa đủ cơ sở để bác bỏ H_0.

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

TH $\sigma_1^2 \neq \sigma_2^2$

- Ta sử dụng thống kê

$$t_0 = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

khi đó t_0 có phân phối Student với bậc tự do *df* được xác định như sau

$$df = \frac{[(s_1^2/n) + (s_2^2/m)]^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} \qquad (3)$$

- Miền bác bỏ trong trương hợp này giống như trong trườn hợp phương sai bằng nhau, chỉ thay bậc tự do *df* cho bởi phương trình (3)

4.3.3 SS2KV - Không biết phương sai, mẫu nhỏ

Bài tập

4.4. Tại một thành phố, khu vực A, người ta chọn ngẫu nhiên 17 sinh viên và cho làm bài kiểm tra để đo chỉ số IQs, thu được trung bình mẫu là 106 và độ lệch tiêu chuẩn bằng 10; tại khu vực B, chỉ số IQs trung bình của một mẫu gồm 14 sinh viên bằng 109 với độ lệch tiêu chuẩn là 7. Giả sử phương sai bằng nhau. Có sự khác biệt về chỉ số IQs của sinh viên ở hai khu vực A và B hay không? Biết$\alpha = 0.02$.

4.5. Hàm lượng thạch tín trong nước càng cao càng có hại cho sức khỏe. Người ta kiểm tra hàm lượng thạch tín ở hai khu vực trung tâm thành phố Biên Hòa và khu vực gần sân bay Biên Hòa. Tại mỗi khu vực, người ta đo ngẫu nhiên hàm lượng thạch tín trong nước ứng với 10 điểm khác nhau. Số liệu cho bởi bảng thống kê bên dưới

Trung tâm TP	3	7	25	10	15	6	12	25	15	7
KV gần sân bay	48	44	40	38	33	21	20	12	1	18

Với $\alpha = 0.05$, hãy kiểm tra xem có sự khác biệt về hàm lượng thạch tín ở hai khu vực này.

4.3.4 So sánh hai tỷ lệ

- Khảo sát những phần tử thỏa một tính chất A nào đó trên hai tổng thể độc lập với tỷ lệ tương ứng là p_1 và p_2; từ hai tổng thể chọn ra hai mẫu với cỡ lần lượt là n và m. Gọi X và Y là số phần tử thỏa tính chất A trong mẫu 1 và mẫu 2. Khi đó, ta có $X \sim B(n, p_1)$ và $Y \sim B(m, p_2)$.
- Bài toán: so sánh hai tỷ lệ p_1 và p_2.
- Bài toán kiểm định giả thuyết, với mức ý nghĩa α gồm các trường hợp sau:

$$(a)\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 \neq D_0 \end{cases} \quad (b)\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 < D_0 \end{cases} \quad (c)\begin{cases} H_0 : p_1 - p_2 = D_0 \\ H_1 : p_1 - p_2 > D_0 \end{cases}$$

- Các giả định
  - Hai mẫu độc lập,
  - Cỡ mẫu lớn, $np_1 > 5; n(1-p_1) > 5$ và $mp_2 > 5; m(1-p_2) > 5$

4.3.4 So sánh hai tỷ lệ

Các bước kiểm định

1. Phát biểu giả thuyết không và đối thuyết,
2. Xác định mức ý nghĩa α,
3. Tính thống kê kiểm định

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}} \tag{4}$$

với

$$\hat{p}_1 = \frac{X}{n}; \hat{p}_2 = \frac{Y}{m}; \hat{p} = \frac{X+Y}{n+m}$$

nếu H_0 đúng, $Z_0 \sim N(0,1)$.

4.3.4 So sánh hai tỷ lệ

4. Xác định miền bác bỏ

Đối thuyết	**Miền bác bỏ**	$p-$**giá trị**
$H_1 : p_1 - p_2 \neq D_0$	$\lvert z_0 \rvert > z_{1-\alpha/2}$	$p = 2[1 - \Phi(\lvert z_0 \rvert)]$
$H_1 : p_1 - p_2 < D_0$	$z_0 < -z_{1-\alpha}$	$p = \Phi(z_0)$
$H_1 : p_1 - p_2 > D_0$	$z_0 > z_{1-\alpha}$	$p = 1 - \Phi(z_0)$

5. Kết luận: Nếu bác bỏ H_0, ta kết luận H_1 đúng với $(1-\alpha)100\%$ độ tin cậy. Ngược lại ta kết luận chưa đủ cơ sở để bác bỏ H_0 với α cho trước.

4.3.4 So sánh hai tỷ lệ

Ví dụ 4.10.

Một công ty sản xuất thuốc cần kiểm tra một loại thuốc có tác dụng làm giảm việc xuất hiện cơn đau ngực ở các bệnh nhân. Công ty thực hiện thí nghiệm trên 400 người, chia làm hai nhóm: nhóm 1 gồm 200 được uống thuốc và nhóm 2 gồm 200 người được uống giả dược. Theo dõi thấy ở nhóm 1 có 8 người lên cơn đau ngực và nhóm 2 có 25 người lên cơn đau ngực. Với mức ý nghĩa $\alpha = 0.05$, hãy cho kết luận về hiệu quả của thuốc mới sản xuất.

4.4.1 So sánh hai kỳ vọng hai mẫu không độc lập

- Khi hai mẫu không độc lập thi mỗi giá trị quan trắc được trong một mẫu có mối liên hệ tương ứng với một giá trị quan trắc ở mẫu thứ hai. Như vậy, ta có thể ghép cặp ứng giá trị trong hai mẫu với nhau.
- Việc ghép cặp là kết quả của việc:
  - quan trắc giá trị trước và sau khi thực hiện 1 thí nghiệm. Chẳng hạn như đo trọng lượng trước và sau khi thực hiện một chế độ **u**an kiêng.
  - so sánh cùng một đặc tính.
  - thí nghiệm trên cùng 1 địa điểm.
  - thí nghiệm với cùng thời gian.
- Xét (X_{1i}, X_{2i}), với $i = 1, 2, \ldots, n$ là tập gồm n cặp gái trị quan trắc với giả sử rằng kỳ vọng phương sai của hai tổng thể đại diện bởi X_1 và X_2 độc lập lần lượt là μ_1 và σ_1^2; μ_2 và σ_2^2. Định nghĩa độ sai khác giữa mỗi cặp trong tập hợp các giá trị quan trắc là

$$D_i = X_{1i} - X_{2i},\ i = 1, 2, \ldots, n$$

4.4.2 Kiểm định giả thuyết về phân phối

Khảo sát biến ngẫu nhiên X liên quan đến một tổng thể có phân phối chưa biết. Cần kiểm định xem phân phối của tổng thể có phải là $F(x, \theta)$ hay không? Ví dụ, ta cần kiểm định phân phối của tổng thể đang xét là phân phối chuẩn.

Các bước kiểm định:

1. Chọn mẫu ngẫu nhiên cỡ $n : (X_1, \ldots, X_n)$. Chia miền giá trị của các biến ngẫu nhiên X_i thành k khoảng không trùng nhau $I_1, I_2, \ldots, I_k$.

2. Gọi O_j là số các giá trị mẫu nằm trong khoảng I_j, $j = 1, 2, ..., k$

3. Phát biểu giả thuyết $H_0 : X$ tuân theo luật phân phối $F(x, \theta)$. Khi đó, tính $p_j = \mathbb{P}(X \in I_j)$.

4. ...

5. ...

4.4.3 Kiểm định giả thuyết về tính độc lập

Giả sử mỗi phần tử trong một tổng thể có thể được phân loại theo hai đặc tính khác nhau, gọi là đặc tính X và đặc tính Y. X có r giá trị và Y có s giá trị. Gọi

$$P_{ij} = \mathbb{P}(X = x_i, Y = y_j) \qquad \text{với } i = 1, \ldots, r, j = 1, \ldots, s.$$

Đặt

$$p_i = \mathbb{P}(X = x_i) = \sum_{j=1}^{s} P_{ij}, \qquad \text{và} \qquad q_j = \mathbb{P}(Y = y_j) = \sum_{i=1}^{r} P_{ij},$$

Ta cần kiểm định xem X có độc lập với Y hay không?
Phát biểu gải thuyết

$$H_0 : P_{ij} = p_i q_j \ \forall i = 1, \ldots, r; j = 1, \ldots, s$$

và đối thuyết

$$H_1 : \exists (i,j) \text{ sao cho } P_{ij} \neq p_i q_j$$